Physics Data Science And Machine Learning Laboratory

Physics 477/677 • Spring 2021 • Christopher Tunnell

Contents

1 Key course information **1**
- 1.1 Course time . . . 1
- 1.2 Location . . . 1
- 1.3 Instructor . . . 1
- 1.4 Teaching Assistants . . . 1
- 1.5 Prerequisites . . . 2

2 Course Overview **2**

3 Course Objectives & Outcomes **2**

4 Team & Project Placement **2**

5 Lectures **3**

6 Coursework **3**

7 Course Expectations **3**

8 Grading rubric **3**
- 8.1 Individual Assessment . . . 4
- 8.2 Weekly Check-ins . . . 4
- 8.3 Presentations . . . 4
- 8.4 Reports . . . 5
- 8.5 Software . . . 6

9 Course Schedule **7**

1 Key course information

Even though similar courses have been done at Rice, this will be the first such course intending to include physics students, which means that the course timeline will be tweaked during the semester in discussion with the students.

1.1 Course time

Wednesday and Fridays at 2:45 PM – 4:05 PM.

1.2 Location

The course is fully online. The Zoom information is in Canvas.

1.3 Instructor

Christopher Tunnell (tunnell@rice.edu)

1.4 Teaching Assistants

None at the moment.

1.5 Prerequisites

This is a physics course rather than a programming course, however it entails significant amounts of programming and reading background literature to understand physics problems. Therefore, you will be assessed on the physical insights that you are able to gain in open physics datasets using computational techniques (though this requires programming). Therefore, some experience in Python is encouraged. It is recommended that students are either:

1. First or second year graduate students who have engaged in research
2. Physics and astronomy seniors engaged in a senior thesis
3. Computational physics students

However, it is advised that you speak with the instructor if you are unclear.

2 Course Overview

:
Modern physics research has been fundamentally changed by the advancement of computing power in the previous decades. In many physics courses, we use mathematics to deeply explore physical phenomena. In this course, we will use physics, statistics, and data analysis software to also gain deep physical insight. Depending on the backgrounds of the students, the difficulty of the computational techniques or physical problems will be adjusted.

In this project-based course, student teams will complete semester-long data science, analysis, and/or machine learning projects based on physics datasets that the students will choose.In practice, the course is similar to summer research projects, but as a group project and with a strong emphasis on developing the core skills needed for modern physics research. This course is adapted from a course developed for the popular D2K program in Engineering. Each student team will have a Rice technical mentor, where this semester it will be the instructor who will help with providing technical training on needed techniques. Student teams are required to meet weekly outside of class. During the semester, teams will work towards project milestones and present their progress in interim reports, oral presentations, and software checks. At the end of the semester, each team will produce a final project report with accompanying software and have the opportunity to present their work in the D2K Showcase. This course will also have a lecture component at the beginning of the semester where students will learn how to design a data science pipeline and best practices in data science, with emphasis placed on computational reproducibility and data exploration. As there will be variation in the technical skills within the groups, we will be learning the necessary computational methods through a mix of short topical lectures and self study.

3 Course Objectives & Outcomes

In this course, students will engage in experiential learning where they learn best practices in physics data analysis by working on real-world projects. Students will learn how to wrangle, explore, and model (incl. potentially machine learning) their data as well as how to design a reproducible data pipeline and validate their findings. Through these projects, students will also learn team-work, leadership, project management, and communication skills. Students completing this course should be able to (i) successfully complete a physics-oriented data science research or analysis project, (ii) successfully devise and employ a computationally and scientifically reproducible data analysis pipeline, (iii) work collaboratively on a team to complete a data science project, write reports, and give presentations, and (iv) effectively communicate their findings to both a general and a technical audience.

4 Team & Project Placement

Teams will consist of ideally five students. Students will be selected for specific projects based on their background and skill set as well as their project preferences. The first two weeks of class center around defining the teams and projects.

5 Lectures

At the beginning of the semester, there will be a small number of lectures on the data science pipeline, how to design a pipeline, technology for data science, and data science reproducibility (see course schedule below). Lecture time will mainly be time in which the teams are working toward their final deliverable ("the final project") where the instructor does check ins with each time. Each team has to prepare for each check-in.

Per request, we can have lectures on specific topics with some notice. However, this is mainly a 'lab' course where lectures will only take up a small amount of class time, where extra recorded lectures will be provided if people want to dive deeper into a topic.

6 Coursework

The major work of this course will be to conduct a physics data-analysis project. Each team will present their work through a final report with accompanying software due at the end of the semester. Throughout the semester, shorter interim reports, presentations, and software checks will be due to give each team feedback and guidance on how best to prepare their final report and presentation. Further details and instructions on the reports, presentations, and software checks as well as grading rubrics are given in a separate document.

7 Course Expectations

Students are expected to work at least 7 hours a week on their project outside of class. Each week outside of class, teams are required to have one meeting with their mentors and at least one separate meeting with just team members. Attendance at team meetings and lectures is required. Additionally, each team member is expected to contribute to all aspects of the data science project including technical data science contributions, coding, and preparing the written report and presentations.

Team Participation: Attendance and team participation will be evaluated through peer evaluation surveys at CATME.org. Students will need to register on this site to access the peer evaluations.

Computing Resources:
Many cloud providers provide free access, especially for students. I recommend that you use Google Colab, which is completely free, for interactive analysis. Additionally, Amazon Web Services, Google Cloud, and Microsoft Azure all provide free credits for students.

Normally, each team is allocated resources through Amazon Web Services (AWS) of $500 of computing credits. However, many cloud providers provide free credits to students (e.g. Google cloud).

Special considerations: Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class. Any issues arising from the pandemic or the remote working situation should also be communicated to the instructor, and will be address on a case by case basis.

Syllabus changes: This syllabus is subject to change with reasonable advance notice by the instructor. As this is the first iteration of a physics version of this course, these changes would be larger than by a normal course. Changes to the course would be to improve this course's effectiveness for the student, so any issues arising from changes should be communicated to the instructor.

8 Grading rubric

Grading Overview:
Individual: 20%
- Peer evaluations: 10%
- Individual contributions: 10%

Team: 80%
- Project Ideas and Proposal: 10%

Initial Evaluations: 10%
Interim Evaluations: 10%
Final Report: 20%
Final Presentation: 15%
Final Software: 15%

Work will be assessed on whether reasonable progress has been made given the expected 7 hour per week work load for this course. As with all physics courses, how you arrived at the answer is more important than the answer. Therefore, even if you single-handedly win the Nobel Prize for discovering some new physical phenomena hidden in some dataset, you will not be successful in the course unless you work with your team toward developing a computationally and scientifically reproducible data-analysis pipeline.

8.1 Individual Assessment

Individuals will be evaluated through two mechanisms: peer evaluation surveys and assessment of individual contributions. First, we will use peer evaluation surveys at CATME.org; students will need to register on this site to access the peer evaluations and can review the evaluation criteria there as well. Second, individual contributions to the team project are assessed via self and team reporting as well as through Teaching Assistants and instructor evaluations. To ensure that individual contributions are accurately assessed, each teammate should clearly delineate their individual contributions to the reports (acknowledgments), in software (via documentation), and in presentations (acknowledgments).

8.2 Weekly Check-ins

During the mentoring stage of the semester, class time will be used for weekly check-ins with Teaching Assistants and the instructor. Weekly check-ins will be graded on a check, check-plus, or check-minus basis and are mainly intended to give feedback and guidance to the team. Each team should prepare some short slides (nothing fancy) for the weekly check-in that covers updates on progress and any new results, challenges and questions that arose, and finally a plan on next steps.

8.3 Presentations

Final Presentations:
Final presentations will include an oral presentation to the class and also a team poster presentation at the D2K Showcase on Dec 15 from 5pm - 7:30pm. A subset of teammates can present orally, with the other teammates presenting the poster. Posters presentations will be judged with prizes given for the top teams. Both the oral and poster presentation will also be graded.

Interim Presentations:
Throughout the semester, teams will give short interim presentations with feedback from the instructor and Teaching Assistants. Each interim presentation should be 5 minutes in length and at least two teammates must present. Over the course of the semester, each teammate must have presented orally at least two times.

Presentations Grading Rubric:

1. Organization: 20%
    - Are the slides / poster well organized and presented in a logical and coherent manner? Are they structured in a way that communicates the problem and results effectively?
2. Oral Communication: 20%
    - Is the oral presentation clear and engaging? Does the presenter demonstrate a clear mastery of the material presented? Are they able to answer follow-up questions effectively?
3. General Audience Communication (Background, Objective & Significance): 20%

- Is the project background, significance and objectives clearly presented, even for a general (lay) audience?

4. Technical Content Communication: 20%

   - Is all technical content communicated clearly and at a level appropriate for the audience?

5. Visual Communication: 20%

   - Are the slides / poster visually engaging? Are graphics easy to understand and interpret, even for a general audience? Are graphics helpful for understanding the data and the results?

8.4 Reports

The final report is due to the instructor and Teaching Assistants at the end of the semester. But, sections of the report / interim reports are due throughout the semester to provide guidance and feedback on team progress.

Report Sections:
Loosely, the report should consist of the following sections. Note that the precise organization / sections of the report are up to the team, but each report should at least contain this content:

1. Introduction:

   - The introduction should be written for a general audience and should outline the problem and its significance.
   - Background: Include background research on the problem / problem area with a literature review.
   - Objectives: State the project objectives clearly.
   - Data Science Pipeline: Give an overview or outline of your data science pipeline design to address the objectives.

2. Data & Data Exploration:

   - Data Description: Fully describe all data used for the project.
   - Data Wrangling: Describe any wrangling done to format the data for analysis.
   - Feature Engineering / Feature Filtering: If any features were engineered or featured filtered, describe the methods used, the resulting features, and why.
   - Exploratory Data Analysis & Visualization: Include a full exploratory data analysis and visual summaries of the data.

3. Methods & Models:

   - Fully describe all data science techniques and models used in the project.

4. Results:

   - Include all project results and explain how to interpret the results in a manner appropriate for a general audience.

5. Conclusion:

   - Conclude by discussing the implications and potential future work related to your project.

Report Grading Rubric:

1. Communication: 30%

   - Is the written communication professional (e.g. no spelling or grammar mistakes)? (Note that ANY spelling or grammar mistake will automatically result in a 10% reduction in the grade.)
   - Visualizations. Are the graphics clear, engaging, informative, and appropriate to convey the results?
   - Is the report well-organized and clear?

- Is the introduction, interpretation of results, and conclusions communicated clearly for a general audience?

2. Significance: 10%
   - What is the significance and impact of this work in both the applied discipline and in data science?
3. Innovation & Creativity: 10%
   - Did the team employ creative or innovative solutions? Apply data science techniques in a novel way? Or develop or extend new data science techniques?
4. Technical Approach: 35%
   - Are sound data science techniques used at each step along the data analysis pipeline? Are techniques and pipeline choices reasonable and justified?
5. Data Scientific Reproducibility: 15%
   - Are best practices and techniques used to validate all data analysis findings?
   - Are all results likely to be scientifically reproducible?

8.5 Software

At the end of the semester, all code and software used in the data science project must be turned in. All software must follow computational reproducibility requirements and use best coding practices. There will be interim software checks throughout the semester to provide each team with feedback and guidance (see separate schedule document).

Software Grading Rubric:

1. Computational Reproducibility: 25%
   - Can all the main results (e.g. figures, tables, etc.) in the submitted report be reproduced using the team's code?
2. Reusability / Data Science Reproducibility: 25%
   - Are best practices in data scientific reproducibility followed?
   - Is the software re-usable on the same or a similar data set?
3. Readability / Organization: 20%
   - Is the code well-organized, modular, and easy to follow?
4. Documentation: 20%
   - Is the code well documented with a clear README file?
5. Efficiency: 10%
   - Does the software use efficient computational techniques to achieve the project objectives?

9 Course Schedule

We meet 26 times this semester across 14 weeks. Please note that this semester is special as it has Spring "Sprinkle" Days as there is no Spring Break this year. Assignment deadlines are indicative, where one should consult Canvas for the latest assignment information.

Date	**Class Time**	**Coursework Due**
January 27, Wed	Welcome, Course, and Idea Generation!	
January 29, Fri	Pick Clusters, Develop 5 ideas	Idea Generation and Sorting, Some setup
February 3, Wed	Establish teams and focus on 3 ideas	Sort ideas from clusters, develop ideas
February 5, Fri	Feedback discussion	Data Set and Project Proposal Ideas Due
February 10, Wed	Check in and/or topical lecture	
February 12, Fri	Initial Feedback discussion	Single Project Proposal Due
February 17, Wed	*No lecture: Spring "Sprinkle" Day*	
February 19, Fri	Discuss detailed feedback Project Proposal	
February 24, Wed	Presentations	Initial Presentations (Pipeline Plan)
February 26, Fri	Software check	Initial Software Check
March 3, Wed	Check in and/or topical lecture	Initial Report (intro + background + objectives + data description + pipeline design plan)
March 5, Fri	Discuss detailed feedback Initial Report	Midterm CATME Eval Due
March 10, Wed	Initial Presentations	Initial Presentations (Intro, Background, Objectives, Pipeline Plan, Data, Data Wrangling, beginning Data Exploration)
March 12, Fri	Check in and/or topical lecture	
March 17, Wed	Check in and/or topical lecture	
March 19, Fri	Check in and/or topical lecture	
March 24, Wed	Software check	Midtern Software Check
March 26, Fri	*No lecture: Spring "Sprinkle" Day*	
March 31, Wed	Check in and/or topical lecture	Midterm Report Due
April 2, Fri	Discussed detailed feedback Midterm report	Midterm CATME Eval Due
April 7, Wed	Midterm presentations	Midterm Presentation
April 9, Fri	Check in and/or topical lecture	
April 14, Wed	Check in and/or topical lecture	
April 16, Fri	Check in and/or topical lecture	
April 21, Wed	Check in and/or topical lecture	
April 23, Fri	Check in and/or topical lecture	
April 28, Wed	Final check in	
April 30, Fri	Final Presentations	Final Oral Presentation
May 12, Wed	*No lecture: End of Semester*	Final Report + Final Software Due + Final CATME Eval Due